A Printed PAW Image Database of Arabic Language for Document Analysis and Recognition

Bilal Bataineh

Department of Computer Science, Deanship of Preparatory Year, Umm Al-Qura University, Al-Abdiyah, Makkah, Saudi Arabia
E-mail: bmbataineh@uqu.edu.sa

Abstract. Document image analysis and recognition are important topics in the field of artificial intelligence. In this context, the availability of a database with good script samples is an important requirement for machine-learning processes. For Latin and Asian languages many suitable databases exist. However, there is a shortage of databases with Arabic samples. In this work, a new database of printed Arabic text is introduced. The new concept of collecting sub-words (PAWs) instead of words or individual character samples was adopted. These PAWs constitute all words in the Arabic language. The collected database consists of 83,056 images of PAWs extracted from approximately 550,000 different words. Each sample is presented in the database in five font types: Thuluth, Naskh, Andalusi, Typing Machine, and Kufi. In total, the database consists of 415,280 images. Moreover, ground truth information is included with each PAW image to describe its occurrence number, occurrence frequency, positions and the shapes of the characters. This paper presents a statistical analysis of the frequency of each PAW in the Arabic language.

Keywords: *Arabic language; database; document images; information retrieval; OCR; PAWs.*

1 Introduction

Document image analysis and recognition (DIAR) is an important topic in pattern recognition (PR). DIAR consists of many techniques, such as optical character recognition (OCR), document image binarization and enhancement, text segmentation, baseline deduction, document layout analysis, and much more [1-4]. These techniques have been used to develop different applications, such as handwritten character recognition, mail sorting, bank check processing, signature and handwriting verification, historical document processing, information retrieval, and others [1-4].

The availability of databases with a sufficient number of adequate samples is an important requirement for researches in DIAR [5]. Such databases provide standard benchmarks that can be used in evaluating the performance of different DIAR systems [6]. For Latin and Asian languages, many standard databases

Received November 4th, 2016, Revised March 7th, 2017, Accepted for publication July 25th, 2017.
 ISSN: 2337-5779, DOI: 10.5614/ itbj.ict.res.appl.2017.11.2.6

have been collected. However, there is a shortage of DIAR databases for the Arabic language [7-9]. A limited number of databases for some techniques such as OCR are available. However, there are no databases for techniques such as Arabic text segmentation or layout analysis. Also, the existing databases are weak, dated or do not support updating based on recent research.

In this work, a new database for printed Arabic text is introduced. We adopted a new concept aimed at collecting the sub-words (PAWs) that constitute the words used in the Arabic language. This partially solves the segmentation challenge and reduces the number of required samples to cover all words. Moreover, a statistical analysis of PAWs and characters used in the Arabic language is provided in this paper. The proposed database is useful for several DIAR techniques, such as OCR, font recognition, segmentation and information retrieval. This database will be available for researchers upon request.

The rest of this paper is organized as follows. Section 2 reviews the literature. Section 3 describes the method used to collect the samples. Section 4 provides a statistical analysis of all aspects relating to the database. The conclusion follows in Section 5.

2 Literature Review

In this section, a review of the unique features and challenges of the Arabic language are provided. Moreover, the best and most recent databases of the Arabic language are presented.

2.1 Features of the Arabic Language

The Arabic language has unique features that distinguish it from other languages, which have an impact on the development of DIAR applications [10,11]. The main features are:

1. The Arabic alphabet consists of 28 letters, written from right to left.
2. Arabic text has only one case; it does not have capital or lower case letters.
3. Letters are connected to each other to construct sub-words (PAWs) and whole words.
4. Each letter can be written in different shapes depending on its position in a word. There are four possible shapes for each letter based on its position: at the beginning, the middle, or the end of a word, or in an isolated position (Figure 1).
5. Some letters (ا ر و ز د ذ) never occur at the beginning or the middle position. Therefore, the following letter is used as the shape for their beginning position (Figure 1).

6. 19 letters have at least one other letter with a similar shape. One, two or three dots are placed over or under the similar letters to distinguish them (Figure 1).
7. Words in Arabic consist of one or more partial sub-words (PAWs). For example, مبرمج (*programmer*) consists of two PAWs, برنامج (*program*) consists of three PAWs, and ديمقراطيه (*democracy*) consists of four PAWs (three sub-words), and so on.
8. The used font style strongly affects the shape of Arabic text.
9. The number of words in the vocabulary of the Arabic language is the largest of all languages in the world [12,13].
10. The Arabic script is used for several different languages, such as Arabic, Persian, Urdu, and others.

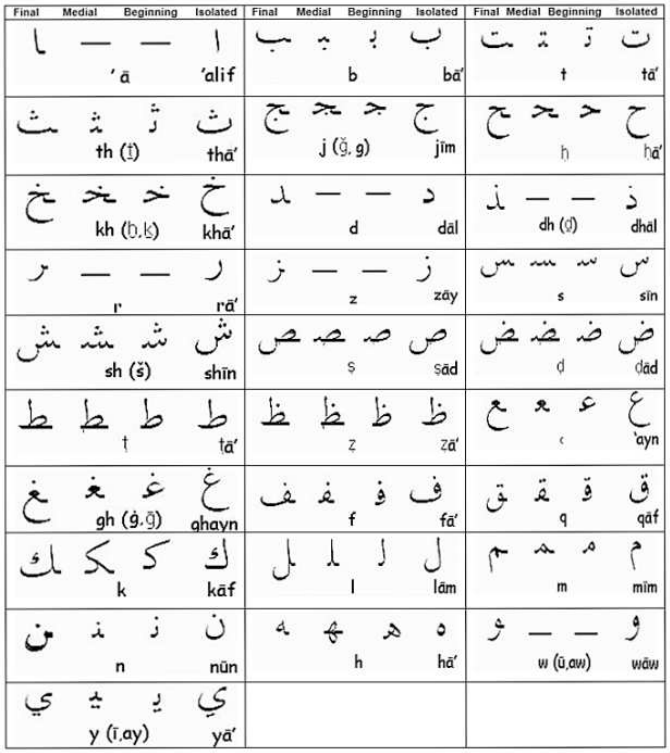

Figure 1 The Arabic alphabet and the shapes for each letter based on its position (beginning, middle, end, or isolated).

2.2 Challenges of Arabic Text DIAR

Based on the above features of Arabic text, there is a set of serious challenges facing DIAR of Arabic text:

1. The segmentation process is an essential challenge in most DIAR applications. In printed Latin and Chinese text, the letters are separate. In Arabic text, however, the letters are connected. This requires adopting a segmentation method that can separate each individual letter. However, determining the shapes, fonts, positions, and occurrence frequency of PAWs is a huge challenge. Up until today, the segmentation of Arabic text needs a great deal of attention, both in the case of printed and handwritten texts [5,14].
2. The large vocabulary of the Arabic language compared with other languages leads to serious problems with regards to the size of databases for machine learning and text recognition.
3. The variance and complexity of the shapes of the letters leads to a reduction of the accuracy rate in the recognition process.
4. Each word may have several similar words with only small differences in shape, e.g. (بصل,يصل, تضل, نضل, تصل,نصل). This negatively affects recognition accuracy.
5. Most languages use the same alphabet, such as English and French both using the Latin alphabet. This greatly helps to improve DIAR applications of those languages. Arabic script requires independent research in developing DIAR applications.

2.3 Database of Arabic Texts

Many databases have been proposed for languages that use the Latin alphabet for use in DIAR research. For example, the IAM database [15,16] is used for handwriting OCR, the RETAS database [7] is used for printed-text OCR, the MNIST database [17] is used for handwritten digit recognition, the NEOCR database [18] is used for natural-image text recognition, the LTP database [19] is used for touching-text segmentation, and the PRImA database [20] is used for printed document layout analysis.

There are a number of databases for Arabic text. Al-Isra [21] is an Arabic handwritten-text images database consisting of 37,000 words, 10,000 digits, 2,500 signatures and 500 sentences. However, this database is not accessible. The IFN/ENIT-database [22] is one of the most famous databases for the Arabic language. It is consist of 26,000 binary word images in 300 dpi that represent Tunisian town/village names. Information about the base line is included. However, this database does not include printed text and it only covers Tunisian town/village names; it ignores other vocabulary of the Arabic language. Moreover, a segmentation process is required before this database can be used for OCR, whereas no methods are available to deal with this accurately. The IfN/Farsi-database [23] only contains Iranian province/city names in Farsi. It is a database of handwritten words consisting of 7,271 binary images of 1,080

Iranian province/city names, collected from 600 writers. This database does not supply more information than the IFN/ENIT-database and the number of samples is smaller. The AHTID/MW [9] database contains images of Arabic handwritten text written by 53 writers. It contains 3,710 text lines with 22,896 words. However, no new contributions were added to this database, while the older databases also comprise handwritten text from more different writers and contain more samples. The AHDB/FTR database [8] consists of Arabic handwritten text images of 497 Libyan city names. In addition to the drawbacks of the previous database, the number of images (497) and writers (five) is very small.

As far as Arabic calligraphy is concerned, [27] presents a database consisting of 69,400 images of Arabic characters using the handwriting of ten calligraphy experts. This dataset introduces a novel feature: the combination of triangle geometry for digital Jawi paleography. Another database for Arabic calligraphy is presented in [28]. This dataset consists of 700 block texture images of the most common types of Arabic calligraphy, where the Kufi, Thuluth, Naskh, Dewani, Andalusi, Ruqqa, and Parisian calligraphy types are represented by 100 images each.

As for printed Arabic text, the ERIM Arabic Document Database is a printed-text database with samples collected from printed documents such as books and magazines [24]. It consists of about 750 pages in several fonts and data qualities but is not accessible for researchers [25]. Another work presents a database containing 6 million samples of printed Arabic words [11]. The samples include words, pieces of words, and pieces of words without diacritics. This database was collected from different resources, such as printed matter, software and websites. However, no standard reference base for the collected words is available.

Based on the above, it can be stated that very few databases for the Arabic language are available. Most of them contain only handwritten text images while printed text is almost completely neglected. Moreover, all the available databases have been proposed for OCR only, while other DIAR applications have been ignored. Also, only the IFN/ENIT handwritten text database is popular and scoped by scientific researches.

The other databases have been created for private research. In addition, there is no central organization to provide a standard reference for the collection and use of Arabic words [10]. Therefore, most of the databases are either very large or extremely small. The quality of a database with separate letters is related to the accuracy of the segmentation method used. Until now, no perfect segmentation

method for Arabic characters exists. On the other hand, the number of samples will be huge if a database is based on words.

3 Proposed Database

In this work, a new principle is proposed, namely to collect all sub-word (PAWs) images of all living words in the Arabic language. To achieve this goal, the following steps were performed.

3.1 Arabic Words Analysis

Referring to the above literature review, the Arabic language has a huge vocabulary, letters are mostly connected, and most words consist of more than one PAW. These three points present the main challenges that face any database for DIAR. In this work, an analytic study was conducted to minimize the impact of these challenges.

The Arabic language uses approximately 11,978 basic roots [26]. This means that a huge number of words can be constructed when the rules of grammar are followed. However, different researchers have reported a large variance in the number of words used in the Arabic language: between 60,000 and 12,000,000 different words. The shapes of words that share the same root are different because of the addition of letters. These letters are connected to the root, producing new shapes. Moreover, most Arabic words consist of more than one part (PAW). Each PAW represents a separate connected component. Figure 2 shows an example of a sentence with four words and containing ten PAWs. A wide space separates different words, while narrow spaces separate the PAWs of each word.

Figure 2 Sentence with four words (separated by wide spaces) with ten PAWs (separated by wide and narrow spaces).

Usually, changes can occur in one PAW while the other PAWs remain unchanged. Moreover, PAWs themselves can also act as letters, whereby different PAWs may be used to construct new words. For example, Table 1 shows 36 different Arabic words consisting of 9 sub-words only. Based on this, if we collect a database consisting of samples of sub-words only, we can cover the whole Arabic vocabulary with a relatively small number of samples.

Table 1 Words made up from nine PAWs.

PAWs	و	سه	ت	سا	سين	ر	مد	ـ	ا
Resulted words	درس	مدرسه	مدارس	مدروسات	مدروسين	دارسات	دارسين	دراسات	دراسه
	لد	رسين	دور	وروار	دوار	در	دار	روس	دروس
	امد	مد	روت	سين	لدت	رسا	وساوسه	ساسه	مدوار
	دسه	مددوا	مدد	مدور	مدا	مداسه	مداسات	ساروا	سار

Based on the above, different Arabic words may consist of a limited number of PAWs (sub-words). Therefore, all Arabic words can be covered by using these PAWs, which effectively leads to a reduction of the number of samples.

3.2 Data Collection

The first step to develop the proposed database was collecting Arabic words in use. Many Arabic words are not used currently and are found only in glossaries. Therefore, the Arabic words were collected from living resources, especially from websites. The world wide web is a huge resource of words in Arabic that are currently in use in all countries. With this resource, we can effectively cover all live Arabic words and determine their occurrence frequency. In total about 550,000 unique words were selected.

Next, each word was separated into PAWs, after which the repeated PAWs were counted. As a result, 83,056 PAW samples were collected, representing all used PAWs in the Arabic language. These samples were stored in PMB image format. Each sample in the database is represented in five font types: Andalusi, Kufi, Naskh, Thuluth, and Typing Machine; each font type is stored in a separate group in the database. In addition, a sub-database consisting of the PAWs of the 1,0000 most-used words in the Arabic language was created. This was done for use in the testing phase of machine-learning research and it can also be useful in researches that prefer not to use all Arabic words.

4 Statistical Analysis

This section presents an analytical and statistical description of the proposed database, whereby the number of images and the used fonts types are highlighted. Then, a detailed statistical analysis of the occurrence frequency of the PAW images is presented. Moreover, we show more information about the ground analysis of each image, including occurrence frequency, meaning, positions and the shapes of the characters for each sub-word image. The proposed database contains the PAWs of the 550,000 most-used words in the

Arabic language. This resulted in 83,056 unique PAW samples. Each sample is represented in five images using the most common font types of printed text in Arabic: Thuluth, Naskh, Andalusi, Typing Machine, and Kufi. The sample images of each font type are grouped into independent sub-databases. In total, the main database consists of 415,280 images of 83,056 different PAW samples.

Initially, 550,000 different words produced the 83,056 PAWs. The statistical analysis of PAW repetition shows that the isolated 'ALIF ا' and the isolated 'WOW و' characters are the most used, i.e. in 177,015 and 71,616 different words respectively. In more detail: about 20 PAW samples contributed to the composition of 10,000 words, and 148 PAWs contributed to the composition of about 10,000 to 1,000 different words. At least 31,395 PAWs are part of more than two different words. Meanwhile, the remainder (about 51,493 PAWs) are used in only one word. In total, the repeated PAWs are used 1,685,202 times, as shown in Figure 3. As a result, all PAWs are repeated 1,736,862 times to produce about 550,000 unique words.

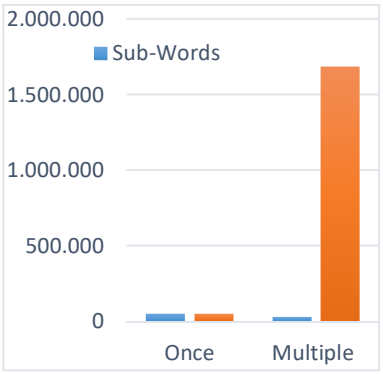

Figure 3 PWAs and the number of words they constitute.

Based on this analysis, most Arabic words consist of more than one PAW. The word images in the database were segmented into individual PAW samples. The following analysis shows that most of the Arabic words consist of two, three or four PAWs. About 181,500 words consist of two PAWs, 180,500 words consist of three PAWs, and 96,000 words consist of four PAWs. The details are shown in Table 2 and Figure 4.

Finally, ground truth information describing each image sample, i.e. occurrence frequency, positions and shapes of the characters in each PAW image are included in the database. Figure 5 shows a snapshot of the ground truth information of random samples (e.g. AR_83042). The first column represents

the image name, the second column represents the text of the image, the third column represents the number of words that use this PAW, the fourth column represents the number of characters in the image, while the last column represents each character and its position in the image. More specifically on character position: 'B' denotes the position at the beginning of a word, 'M' denotes the position in the middle, 'E' denotes the position at the end, and 'I' denotes an isolated position.

Table 2 Number of the words based on number of constituent PAWs.

Nr. of PAWs	1	2	3	4	5	6	7	>=8
Nr. of words	7000	181500	180500	96000	28000	6000	1100	300

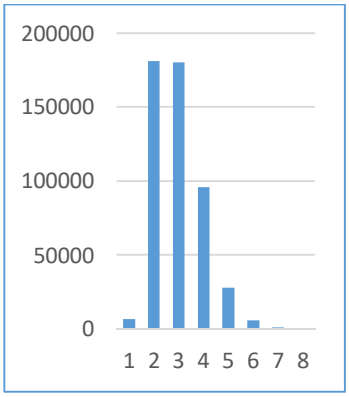

Figure 4 Number of words based on number of constituent PAWs.

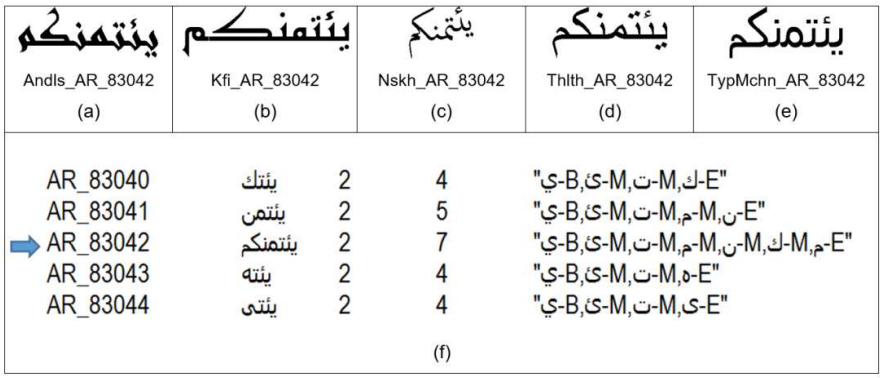

Figure 5 The five images of sample AR_83042 in five font types, (a) Andalusi, (b) Kufi, (c) Naskh, (d) Thuluth, and (e) Typing Machine. Meanwhile, (f) is its ground truth information.

Table 3 and Figure 6 show the analytical results of the number of characters in each PAW image. Most of the Arabic PAWs consist of five, four, and then six characters. Rarely, the PAW images consist of one, two, or more than nine characters. The highest number of characters in a PAW image is eleven, occurring in three cases only.

Table 4 Number of PAW images based on number of constituent letters.

# letters	1	2	3	4	5	6	7	8	9	10	11
# PAWs	38	729	6198	21239	28370	15050	6633	1525	237	34	3

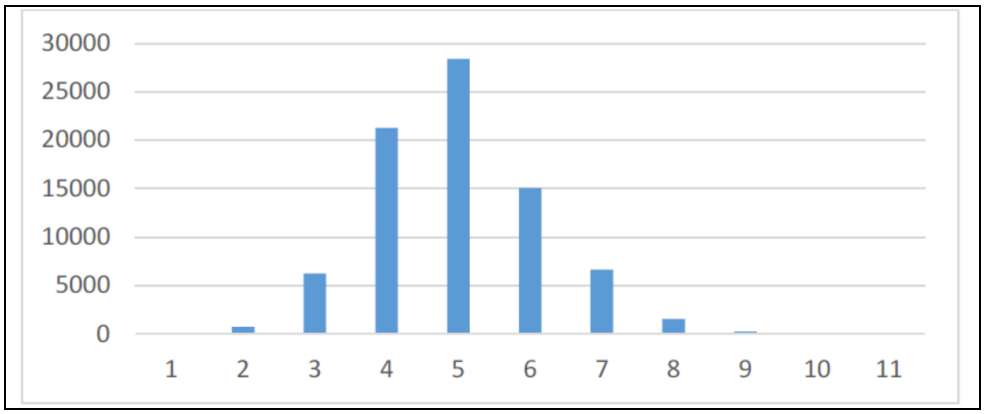

Figure 6 Number of images in the database based on number of constituent characters.

5 Conclusion

In this research, a database for printed Arabic text was collected. This database consists of image samples representing all sub-words (PAWs) that can be used to construct any word in the Arabic language. In total, the proposed database consists of 415,280 images. These images represent 83,056 unique samples of PAWs that can be used to construct approximately 550,000 different words in Arabic.

Each of the collected samples is represented five times in different font types, i.e. Thuluth, Naskh, Andalusi, Typing Machine, and Kufi. Moreover, ground truth information describing each image sample is also included in the database,

i.e. occurrence frequency, positions, and the shapes of the characters for each PAW image. This database is assumed to be useful for several DIAR applications, such as OCR, segmentation and information retrieval.

References

[1] Abu-Ain, T., Abdullah, S.N.H.S., Bataineh, B., Omar, K. & Abu-Ein, A., *A Novel Baseline Detection Method of Handwritten Arabic-script Documents Based on Sub-Words*, Soft Computing Applications and Intelligent Systems, Springer. Shah Alam, pp. 67-77, 2013.

[2] Bataineh, B., Abdullah, S.N.H.S. & Omar, K., *Adaptive Binarization Method for degraded Document Images Based on Surface Contrast Variation*, Pattern Analysis and Applications, pp. 1-14, 2015.

[3] Breuel, T.M., Ul-Hasan, A., Al-Azawi, M.A. & Shafait, F., *High-performance OCR for Printed English and Fraktur using LSTM Networks*, 12th International Conference on Document Analysis and Recognition, IEEE. Washington, DC, 2013.

[4] Breuel, T.M., Ul-Hasan, A., Al-Azawi, M.A. & Shafait, F., *Document Image Quality Assessment Based on Texture Similarity Index*, IEEE Workshop on Document Analysis Systems (DAS), 12th IAPR, Santorini, Greece, 2016.

[5] Bataineh, B., Abdullah, S.N.H.S. & Omar, K., *A Novel Statistical Feature Extraction Method for Textual Images: Optical Font Recognition*, Expert Systems with Applications, **39**(5), pp. 5470-5477, 2012.

[6] Ntirogiannis, K., Gatos, B. & Pratikakis, I, *ICFHR2014 Competition on Handwritten Document Image Binarization (H-DIBCO 2014)*, IEEE 14th International Conference on Frontiers in Handwriting Recognition, Crete Island, Greece, pp. 809-813, 2014.

[7] Yalniz, I.Z. & Manmatha, R, *A Fast Alignment Scheme for Automatic OCR Evaluation of Books*, IEEE International Conference on Document Analysis and Recognition, Beijing, China, 2011.

[8] Ramdan, J., Omar, K., Faidzul, M. & Mady, A., *Arabic Handwriting Data Base for Text Recognition*, Procedia Technology, **11**, pp. 580-584. 2013.

[9] Mezghani, A., Kanoun, S., Khemakhem, M. & El Abed, H., *A Database for Arabic Handwritten Text Image Recognition and Writer Identification*, IEEE International Conference on Frontiers in Handwriting Recognition (ICFHR), Bari, Italy, 2012.

[10] Alginahi, Y.M., *A Survey on Arabic Character Segmentation*, International Journal on Document Analysis and Recognition (IJDAR), **16**(2), pp. 105-126, 2013.

[11] AbdelRaouf, A., Higgins, C.A. & Khalil, M, *A Database for Arabic Printed Character Recognition*, in International Conference Image Analysis and Recognition, Springer, Portugal, 2008.

[12] Al-Fassi, A., Al-Tanji, M. & Al-Ashbili, A.B., *The Conclusion of Vision,* Ministry of Endowments and Islamic Affairs, Morocco, 1963. (Text in Arabic)

[13] Al-Raafiy, M.S., *The History of Arab Etiquette*, Dar al-Kitab al-Arabi, Egypt, 1997. (Text in Arabic)

[14] Gaddour, H., Kanoun, S. & Vincent, N., *A New Method for Arabic Text Detection in Natural Scene Image Based on the Color Homogeneity*, in International Conference on Image and Signal Processing, Trois-Rivières, Canada, Springer, 2016.

[15] Marti, U.V. & Bunke, H., *A Full English Sentence Database for Off-Line Handwriting Recognition*, Proceedings of the Fifth IEEE International Conference on Document Analysis and Recognition, ICDAR'99, Bangalore, India, 1999.

[16] Marti, U.V. & Bunke, H., *The IAM-database: An English Sentence Database for Offline Handwriting Recognition*, International Journal on Document Analysis and Recognition, **5**(1), pp. 39-46, 2002.

[17] LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P., *Gradient-based Learning Applied to Document Recognition*, Proceedings of the IEEE, **86**(11), pp. 2278-2324, 1998.

[18] Nagy, R., Dicker, A. & Meyer-Wegener, K., *NEOCR: A Configurable Database for Natural Image Text Recognition*, in International Workshop on Camera-Based Document Analysis and Recognition, Springer, Beijing, China, 2011.

[19] Kang, L., Doermann, D., Cao, H., Prasad, R. & Natarajan, P., *Local Segmentation of Touching Characters Using Contour Based Shape Decomposition*, 10th IAPR International Workshop on Document Analysis Systems (DAS), IEEE, Queensland, Australia, 2012.

[20] Antonacopoulos, A., Bridson, D., Papadopoulos, C. & Pletschacher, S., *A Realistic Database for Performance Evaluation of Document Layout Analysis*, 10th International Conference on Document Analysis and Recognition, IEEE. Barcelona, Spain, 2009.

[21] Kharma, N., Ahmed, M. & Ward, R, *A new Comprehensive Database of Handwritten Arabic Words, Numbers, and Signatures Used for OCR Testing*, IEEE Canadian Conference in Electrical and Computer Engineering, Alberta, Canada, 1999.

[22] Pechwitz, M., Maddouri, S.S., Märgner, V., Ellouze, N. & Amiri, H., IFN/ENIT-Database of Handwritten Arabic Words, in Proc. of CIFED, Citeseer, **2**, pp. 127-136, 2002.

[23] Mozaffari, S., El Abed, H., Märgner, V., Faez, K. & Amirshahi, A., IfN/Farsi-Database: A Database of Farsi Handwritten City Names, International Conference on Frontiers in Handwriting Recognition, 2008.
[24] Schlosser, S., ERIM Arabic Document Database, Environmental Research Institute of Michigan, 2002.
[25] Doermann, D. & Jaeger, S., *Arabic and Chinese Handwriting Recognition,* Springer-Verlag Berlin Heidelberg, 2006.
[26] Al-Zoubaidy, M. *The Bride Crown from the Jewels of Dictionaries*, Arabic-arabic dictionaries, Dar Al Hedaya, Damaskus, Lebanon, 1965. (Text in Arabic)
[27] Sanusi, M.A., *A Novel Feature from Combinations of Triangle Geometry for Digital Jawi Paleography*, PhD Dissertation, Department of Computer Science, University Kebangsaan Malaysia, 2013.
[28] Bataineh, B., Abdullah, S.N.H.S. & Omar, K., *A Novel Statistical Feature Extraction Method for Textual Images: Optical font recognition*. Expert Systems with Applications, **39**(5), pp. 5470-5477. 2012.